Trust criteria for artificial intelligence in health: normative and epistemic considerations

Kristin Kostick-Quenet ,[1] Benjamin H Lang,[1,2] Jared Smith,[1] Meghan Hurley,[1] Jennifer Blumenthal-Barby[1]

[1]Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, Texas, USA
[2]Department of Philosophy, University of Oxford, Oxford, Oxfordshire, UK

Correspondence to
Dr Kristin Kostick-Quenet, Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, Texas, USA; kristin.kostick@bcm.edu

Received 16 June 2023
Accepted 2 November 2023

To cite: Kostick-Quenet K, Lang BH, Smith J, *et al*. *J Med Ethics* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jme-2023-109338

ABSTRACT

Rapid advancements in artificial intelligence and machine learning (AI/ML) in healthcare raise pressing questions about how much users should trust AI/ML systems, particularly for high stakes clinical decision-making. Ensuring that user trust is properly calibrated to a tool's computational capacities and limitations has both practical and ethical implications, given that overtrust or undertrust can influence over-reliance or under-reliance on algorithmic tools, with significant implications for patient safety and health outcomes. It is, thus, important to better understand how variability in trust criteria across stakeholders, settings, tools and use cases may influence approaches to using AI/ML tools in real settings. As part of a 5-year, multi-institutional Agency for Health Care Research and Quality-funded study, we identify trust criteria for a survival prediction algorithm intended to support clinical decision-making for left ventricular assist device therapy, using semistructured interviews (n=40) with patients and physicians, analysed via thematic analysis. Findings suggest that physicians and patients share similar empirical considerations for trust, which were primarily *epistemic* in nature, focused on accuracy and validity of AI/ML estimates. Trust evaluations considered the nature, integrity and relevance of training data rather than the computational nature of algorithms themselves, suggesting a need to distinguish 'source' from 'functional' explainability. To a lesser extent, trust criteria were also relational (endorsement from others) and sometimes based on personal beliefs and experience. We discuss implications for promoting appropriate and responsible trust calibration for clinical decision-making use AI/ML.

INTRODUCTION

Rapid advancements in artificial intelligence and machine learning (AI/ML) in healthcare have raised pressing questions about how much users should trust AI/ML systems, particularly for high stakes clinical decision-making. To date, the AI ethics literature has highlighted concerns about accuracy, bias, transparency and trustworthiness of AI/ML systems, issues which are now globally recognised as paramount for responsible AI/ML governance. The USA and EU have proposed policy frameworks to advance trustworthy AI, both separately[1 2] and together in their recent 'TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management' (December 2022).[3] These high-level efforts to develop responsible AI/ML are crucial; however, they must also be informed by empirical insights into trustworthiness criteria from users and stakeholders of real AI/ML tools in real healthcare settings as well as normative reflection around the ethical and practical impacts that may result from placing trust in these systems.

The pillars of trustworthiness—for example, those identified and defined by the National Institute of Standards and Technology[2]—are likely to vary in importance across different tools, use settings and end-user dynamics.[4] For example, a risk predictor for cardiovascular disease may raise different ethical and trust considerations than a radiomics tool for cancer detection. Identifying which elements of trustworthiness to prioritise for which tools in which environments should be an ongoing priority. Furthermore, different stakeholders may prioritise different trust criteria, with little agreement on which criteria constitute a 'ground truth' for evaluating AI/ML trustworthiness. Examining variation in trust criteria among stakeholders provides crucial insights into how these tools are likely to be used in real settings (eg, under-reliance vs over-reliance), and the degree to which these usage patterns align with emerging recommendations for high stakes AI/ML use. This effort draws a distinction between how much stakeholders *do* versus *should* trust these systems, based on emerging evidence about their performance capacities and potential impacts. It also highlights the distinction between trust (an attitude) versus reliance (a behaviour) made by Ajzen and Fishbein.[5] While the latter is often considered a behavioural extension of the former, trust is not strictly determinative of behaviour, as other factors can bypass trust to directly influence reliance on machines. Organisation, cultural and environmental factors such as workload and need for multitasking (eg, in high pressure or fast-paced environments) can increase reliance on machines by necessity,[6 7] even if trust is low. Furthermore, psychosocial factors like situational awareness,[8] self-confidence of the operator and level of task expertise[9] can impact users' reliance on machines. Thus, the level of reliance exhibited by peoples' behaviours may not accurately reflect their underlying trust, nor represent optimal ways of interacting with machines.

Our study is among the first to report on stakeholders' subjective perspectives on considerations for trust and trustworthiness of AI/ML-based models for risk prediction for a high-stakes medical procedure: implantation of a Left Ventricular Assist Device (LVAD) for patients with advanced heart failure. Identifying trust criteria for an algorithm with such significant survival and mortality implications in contexts of high outcome uncertainty may offer broader normative insights about how physician and patients should (and should not) integrate

Table 1 Sample demographics

	Patients		Coordinators		Physicians		
	% Total		% Total		% Total		Total
n =	18	47%	6	16%	14	37%	n=38
Male	11	29%	1	3%	12	32%	64%
Female	7	18%	5	13%	2	5%	36%
White	8	21%	4	11%	8	21%	53%
AA/Black	10	26%	2	5%	0	0%	31%
Asian	0	0%	0	0%	6	16%	16%
Other	0	0%	0	0%	0	0%	0%
Pre-LVAD	1	3%	--	--	--	--	3%
Post-LVAD	17	45%	--	--	--	--	45%

LVAD, left ventricular assist device.

AI/ML risk prediction tools into high stakes clinical decision making than has been established in previous research.

BACKGROUND

The particular AI/ML tool examined in our research with stakeholders is a risk prediction calculator designed to predict survival outcomes for patients with advanced heart failure after receiving a LVAD. At the time of interviewing, the calculator was in the early stages of development and intended to use Bayesian machine learning to predict personalised patient outcomes related to both survival and adverse events postimplantation, and to be integrated into a validated decision support platform[10–12] to inform shared decision-making about LVAD among clinicians, patients and caregivers. We conducted semistructured interviews as part of an ongoing (5 year), multisite study to explore stakeholder attitudes towards integrating ML-based personalised risk (PR) estimates into clinical decision-making for patients considering an LVAD. Here, we present results specifically focused on stakeholders' attitudes towards the acceptability and perceived trustworthiness of such a calculator.

METHODS

We conducted 40 in-depth interviews with stakeholders recruited (January–November 2021) from six academic medical centres, including physicians with LVAD expertise, nurse coordinators, patients and their caregivers in clinical decisions about LVADs (table 1). Our interview guide was developed by our team of ethicists and qualitative research methodologists based on issues raised in the clinical and ethics literature related to risk communication, the use of AI/ML tools in healthcare as well as through our 7 years of experience working with our stakeholder populations.[10–12] We piloted our guide with two participants from each stakeholder type. Questions relevant to this paper explored perceived trustworthiness, accuracy and relevance of PR scores for decision-making. The study was approved by the Institutional Review Board at Baylor College of Medicine. In line with our IRB's approved waiver of documented written consent, all respondents offered verbal informed consent to participate. The number of interviews was not predetermined but rather arrived at using a criterion of diminishing informational returns from each subsequent interview (ie, 'saturation'). Of the 40 interviews conducted, we left out two caregivers from this analysis because we did not feel we had enough data from this group to make confident conclusions about their perspectives. This omission resulted in a total of 38 interviews.

Participant sampling

Interview participants were recruited from six partnering sites. We identified potential participants through physicians and nurse coordinators on our advisory panel. Our primary sampling criteria for physicians and nurse coordinators were that they are actively involved at providing patient education and/or care to patient candidates for LVAD. Inclusion criteria for patients and caregivers is that the patient was actively engaged in or had already engaged in decision-making regarding LVAD treatment. While stakeholder demographics were not part of our sampling criteria, we did document demographics such as age, gender, ethnicity and LVAD status, both for reporting requirements and to explore any systematic variation by demographics in stakeholders' perspectives. We contacted potential candidates by email or phone by a research coordinator (BHL) and attempted to include perspectives from patients both pre-LVAD and post-LVAD. Interviews lasted on average 40 min (±7.4 min) and were conducted via phone or videoconference by members of the research team trained in qualitative interviewing by a medical anthropologist (KK-Q). Patients and caregivers received compensation for their time.

Data analysis

Interviews were audio-recorded, transcribed verbatim, and coded and analysed using MAXQDA 2020.[13] Team members (JB-B, KK-Q, BHL) collaboratively developed a codebook and issues related to perceived acceptability, trust and accuracy were explored across code outputs. Four research assistants were trained in qualitative analysis by a medical anthropologist (KK-Q) and engaged in two rounds of preliminary coding to reduce interpretive variation in coding styles and applications. Each transcript was coded by merging work from at least two coders. We examined intercoder reliability in two preliminary phases of analysis during which all coders coded the same interviews (n=4 transcripts, with two transcripts coded in each round). We examined intercoder reliability represented as the percentage of times any two coders applied the same code to the same text within and across documents with at least 25% overlap of the selected text segment(s), resulting in a interreliability matrix of 4×4 coders (KK-Q, BHL, MH, JS). After the first preliminary coding phase, we collaboratively reviewed discrepancies in coding applications with a goal of achieving greater consensus in interpreting text segments in line with our code concepts. We then engaged in another coding round and repeated this step, resulting in fewer interpretive discrepancies. To reduce even further the potential for interpretive bias in coding applications, we selected two coder *pairs* to code the remaining transcripts moving forward. We selected coder pairs by systematically matching coders with the lowest intercoder reliability (ie, those with maximally different interpretations), resulting in two coder pairs for each document going forward, allowing us to halve the total number of transcripts between these coder pairs. Thematic Content Analysis[14] was used to inductively identify themes by progressively abstracting relevant quotes, a process that entails reading every quotation to which a given code was attributed, paraphrasing each quotation (primary abstraction) and further identifying which constructs were addressed by each quotation (secondary abstraction). To ensure credibility and dependability of our findings, all abstractions were validated by at least one other member of the research team before calculating thematic frequencies to characterise stakeholders' responses. In rare cases where abstractions reflected different interpretations, members of our research team met to reach consensus.

FINDINGS

Overall, physicians and patients tended to share similar considerations and criteria for trust, though they discussed them via different lexicons. In order of prevalence, trust considerations for both stakeholder groups were primarily *epistemic* in nature, focused on the accuracy and validity of PR scores. However, we also identified trust criteria that were *relational* as well *personal belief-based*. We elaborate on these findings below.

Epistemic considerations for trust

Accuracy

For both stakeholder types, trust was overwhelmingly discussed in terms of the perceived accuracy of algorithmic outputs (see table 2). For physicians, in particular, accuracy was often discussed in statistical terms, for example, via measures of positive predictive value, including area under the curve statistics. Physicians disagreed on how much accuracy is necessary, with some citing requirements of ~70% and others demanding upwards of ~90% accuracy. Some physicians noted that they demand higher levels of accuracy for estimates intended to inform choices about pursuing an intervention with life-or-death implications versus those intended to predict postoperative adverse events and manage postoperative care. Patients, on the other hand, discussed accuracy in simpler terms of whether an algorithm was 'right' or 'wrong'. Both physicians and patients stressed the need for models to be prospectively validated for accuracy.

Validity

Both physicians and patients also described how their trust is contingent on the relevance of training data sets, validation cohorts (eg, emphasising sufficient size and heterogeneity) and target outcomes. For example, physicians pointed out that the relevance and accuracy of estimates are outcome-specific, with some outcomes easier to predict than others (eg, gastrointestinal bleeding vs postimplant survival, respectively). Furthermore, they argued that trustworthy estimates must account for variation across clinical sites, surgical teams and specific patient management approaches, as well as be device-specific, given that models trained on outdated devices will not perform as well for state-of-the-art devices. Physicians also noted the limited trustworthiness of models trained on data sets in which all patients received an LVAD, citing the equal value and relevance of learning from non-candidate and decliner outcomes.

Both physicians and patients also underscored challenges of estimating outcomes due to significant patient heterogeneity. Some physicians cited unanticipated intraoperative events, while certain patients cited individual differences (eg, '…everybody has their own risk factor' (patient 09) and '…estimates are unique to the individual' (patient 03)). Patients in particular argued that algorithmic estimates cannot account for patient agency and determination to change health-related behaviours and influence outcomes.

Finally, both physicians and patients said that trustworthiness of algorithmic estimates is limited by their inability to account for the full range of factors impacting outcomes, including chance and/or accidental trauma (eg, to driveline).

Many physicians and patients alike shared that their trust in algorithmic estimates is contingent on the degree to which systems can reveal the explanatory power of certain input variables (see table 3). Physicians pointed out that explainability becomes particularly important when estimates differ from a physicians' own clinical intuitions or evaluations. Patients similarly wanted to know which sources of information a model uses to derive an estimate and expressed concern over whether these data sources were relevant or sufficient to calculate a trustworthy estimate.

Some physicians, on the other hand, deemphasised the importance of explainability, particularly in relation to other factors cited as critical for algorithmic trust (eg, validation in relevant patient cohorts; endorsement by peers). However, physicians noted that explainability may be important for communicating risk effectively to patients.

Relational trust considerations

Both physicians and patients explained that they would be more likely to trust algorithmic estimates that were endorsed by members of the medical community who are themselves perceived as reputable and trustworthy (see table 4). Physicians, in particular, said they would look for validation studies published in reputable, peer-reviewed journals and whether the model was highly regarded or endorsed by reputable colleagues. Patients expressed that they would likewise be more likely to trust a model if it was endorsed by their clinical team.

Personal belief-based trust considerations

A minority of physicians and patients described how their trust in algorithmic estimates is influenced by their personal beliefs (eg, in their own vs an algorithm's predictive abilities as well as personal beliefs about prediction in general). For example, certain physicians expressed a view that algorithms cannot (yet) approximate physicians' expertise, suggesting a broader belief in the inferiority of AI/ML tools related to human experts. Certain patients also suggested that they place greater trust in faith and prayer than in algorithmic estimates, and others emphasised the inability of algorithmic estimates to account for the power of hope and optimism in determining outcomes. Still others expressed distrust in algorithms but could not explain why.

DISCUSSION

Our findings suggest that physicians and patients share a similar set of considerations when deciding whether to trust AI-derived risk estimates. Here, we consider the ethical and practical implications of these considerations for acceptability and uptake of AI/ML algorithms in healthcare. In our study, both patients' and physicians' primary trust needs reflect epistemic considerations (where knowledge comes from) and, to a much lesser extent, relational ones (who else endorses this source of knowledge, similar to what Rempel *et al* referred to as 'network trust'[15]) and personal ones (about the nature of AI-based knowledge and prediction in general). These findings directly challenge a pervasive assumption that patients are willing to blindly trust an algorithmic tool simply because their clinical team endorses it. Instead, patients, just as much as physicians, demonstrated a desire to make reflective decisions about whether to trust AI/ML estimates. Our findings suggest that users *intend* to engage in reflective decision-making about whether to trust AI tools. This finding raises important implications for ethical and regulatory approaches to integrating AI in healthcare, including questions around agency and the 'right to know' whether AI is being used in one's healthcare. Recent regulation issued by the US government concerning citizens' rights around AI, the Blueprint for an AI Bill of Rights[16] raised the 'right to know' as an imperative and our findings support a basic notion that individuals have specific informational needs around the use of AI in their care.

Both physicians and patients primarily wanted to know whether an algorithm's estimates were demonstrably 'right'

Table 2 Epistemic considerations for trust

	PHYSICIANS	PATIENTS
Accuracy	'To me, at the end of the day **I just want a tool that is fairly accurate in terms of predictability**. How it's derived, I feel I place less of an emphasis on this and that **if it works well then it works well.'** (03)	'I can tell you my personal opinion is, I listen and I make up my own decisions cause **I've been told wrong information before**… I could tell you many stories of… information I've been given that was different from three different people. And so **[risk predictions are] not always accurate** and made for your particular situation.' (18)
	'You really need to have RSCs greater than .70-.75, all right?… I think that's sort of a bare statistical minimum if you're going to be telling an individual patient that we can predict what's going to be happening to you.' (01)	
	'I think for survival and RV failure… **I really want to [see a] higher bar**. I think something **[an AUC metric] over .09 would be something that I would have to see.'** (03)	
	'I wouldn't be looking for sensitivity or specificity… **I would actually like to see negative predictive value and positive predictive value for the model performance for the question at hand…** The question is how much of it can you believe?… Of course it needs to be validated in a large cohort.' (05)	
Validity	'I think at the end of the day… **it goes back to how well it's validated**. If you're able to, for example, have a pool of 100 patient cases where this tool is utilized and we see that the predictability or the accuracy of some of these predictions in terms of adverse outcomes is very good, then that to me is **more important than knowing how it's derived or what kind of statistics go into the calculation of these numbers.'** (06)	**'If these numbers are coming from actual studies and patients, I'd be pretty trustworthy on that… I'd be kind of trusting knowing that these are actual numbers, these are actual results of what somebody went through**. Like, you're not trying to hide anything from me and just make it sound good.' (02)
	'You have such a relatively small data set. **You only have a few thousand a year. You don't have 200 000 where you can say that [with certainty**].' (09)	'Let me think about it a little bit. It would depend on what the study expects, right? Like **what's the variation in the study, in the prediction. So, what's the confidence level that you're telling me about….** I would say something like **this could be 40%, 50% of [my] decision because all the inputs are there from a clinical standpoint**. Beyond that, I'll have to look at the support system and some of those other factors.' (12)
Relevance		
Population-specific	**'The major bias is that these are patients that have already been selected for LVADs, right?** So any time you use INTERMACS or UNOS for post-transplant, **what you're looking at is not a general population, it's a population that every center has made a decision about whether this patient is a good candidate for LVAD. So [with this data], I can tell you about folks who end up getting LVAD** but cannot tell you prospectively, what if you decide that patient does not get an LVAD, right?' (14)	
Patient - specific: idiosyncratic	'How's the operation going to go? There are so many factors that can [happen] intraoperatively, and([43])all bets are off.' (09)	**'The accuracy would be kind of iffy, because everybody has their own risk factor**… I really think that the **accuracy is maybe not as high as I would like it to be**, so I'm not really sure how to put that into words.' (09)
		'I'm fully aware that everybody's situation is unique to their own individual set of circumstances. So I don't say, 'Oh my God, **you told me I had a 75% chance that my quality of life would improve.' I know nobody can do that for me… It's going to be unique to the individual.'** (03)
Patient-specific: agency/ behaviour		'If you do what you're supposed to do, I think [the trustability] it's very slim. As far as taking your meds and not doing [consuming] salt and everything that they ask you to do… **you can help prevent it [negative outcomes] by doing the right thing**. But whatever was meant to be is meant to be, first of all. But **if you can help better instead of worsen the process, then that's what you do.'** (07)
Outcome - specific	'The GI bleeding risk is probably something that you can predict very well. **I would probably take into consideration an [algorithm]-provided percent bleeding risk for a patient**… I would say **that would probably exceed and improve what I would kind of eyeball…([44])we're discovering new risk factors to arteriovenous malformations. It's too much in flux. This is not like [you could ask the calculator],** 'Could I get a stat for an LDL of 109?' and then look at 20 000 patients.' (09)	
Site-specific	**'[The data] are not center-specific…surgeon-specific [nor] specific to the post-operative care that the team delivers**. We [at my clinic] have a very… multidisciplinary approach. [When I came here], the one year survival post-transplant rate was 75% for the HeartMate II, which was national average at the time. Then we changed a lot of things in the management of the patients mostly, and we reached 100%, and our average is now low 90 s. **So, and this is the same patients([44])different management.'** (09)	

Continued

Table 2 Continued

Device-specific	'I think **it's hard to compare different LVADs**, and I know they [INTERMACS] include a lot of LVADs, HeartMate 2, 3, HeartWare's all in there, some of probably the older ones. I think when we looked at our institution, you can sometimes see [where the data is from, but]….([44])sometimes, we just don't know. I know they [model developers] try to make the data work… but sometimes, you just don't know.' (08)	
	'Any and all data that I use from INTERMACS have to be specifically for the HeartMate 3.' (09)	
Chance	'It needs to be very clearly laid out to these patients that there are **things that can happen that these models absolutely don't account for**. [With] the LVAD, trauma to the exit site is a one of the number one predictors of driveline infection. And once you get a driveline infection, everything starts to spiral down. **How can you develop a predictive model that's gonna tell you if a patient is going to drop their controller and tug at the driveline exit site – it'd be pretty hard to do, right?'** (01)	'It's [algorithmic risk estimate] just such a generalisation really. Yeah, I could be in the group that lives 36 months based on my heart condition, but **what's to say that my machine won't break in 12 months?** Knock on wood, I've been two and a half years without any issues on my LVAD… but I could go into the hospital tomorrow with my drive line ripped out and hanging by a thread, and they'd have to crack me open again. **There's no telling. Real life. There's value in this [a personalized risk calculator], but… there are so many variables that you don't understand or can't predict.'** (11)

LVAD, left ventricular assist device.

(accuracy), how often (reliability) and for whom (relevance; generalisability). Their evaluations for trust focused more on *where* data come from versus *how* data inputs are computationally processed. While both patients and physicians wanted to know which factors most account for estimate outputs, members of neither group voiced substantive concerns about using opaque 'black box' algorithms nor a fundamental distrust in the computational processes themselves underlying complex machine learning approaches such as deep learning—issues that the normative literature has focused heavily on. Neither stakeholder type professed a need to be able to understand the underlying mathematical calculations (ie, 'look under the hood') by which an algorithm reaches a conclusion, as is often proposed in the literature on AI trustworthiness and explainability.[17–19] Instead, both patients and physicians from our study emphasised a desire to know more about the nature of data sets—rather than algorithms—used to train algorithmic models, in order to gauge relevance of outputs for making inferences about target users or subjects. This finding suggests a need to distinguish between what might be called 'source explainability' versus 'function explainability', that is, a focus on the nature and quality of data sources selected to train an algorithmic

Table 3 Epistemic considerations for trust

		Physicians	Patients
Explainability			
	Variable selection	**'With humility to the computer being able to process so much more than we can, it would be helpful to see where it's coming from, to see and explain it. Maybe you can't explain all of it, so I think it would definitely be helpful to see which variables are driving it [the outputs**]. Obviously, **if I want to put an LVAD in someone, and by all my normal criteria, I think I should, [and] the machine tells me I shouldn't or I should give a lot of pause… [if] I don't know what variables are driving that, it's going to be hard for me to change my decision**… It's going to be hard for me to go against my clinical gut decision, if I can't see why the computer's telling this is higher risk.' (03)	'A lot of **these assessments are not taking things into consideration like diabetes, exercise levels, stuff like that, and those can have a big impact on how long you live**. So that's why I say only about 50% to 75% [accurate], because **there's more to it than just the things that you're looking at for those six rows there**.' (04)
		'You would want to know that it's [the algorithm is] incorporating the kinds of variables that you heuristically turn to and feel are the most important for making a decision.' (07)	**'Where are you getting the information from? Are you doing blood work. What are you doing? Are you just looking at me, looking at my heart? What's going on, how are you guys analyzing it?'** (06)
			'[If] I was presented with, 'You're going to be gone in a few days,' or 'This is going to keep you going,'… **my only question [would be): How did they figure, how would they know? I mean, it's just a pretty rough assessment of your medical history?'** (18)
Explainability (de-emphasised)		**'For me, the black box issue is not a huge issue**. What's important is if it's based on published results, et cetera, and in good patient population. **I know for others it's probably more important. But for me, I find the black box issue less of an impediment**.' (14)	
		'I don't think [the black box issue] that's such a big deal. I think those of us that are tech savvy, most cardiologists are, understand the added value of applying machine learning tech, adequately enough, that you don't need to really explain the black box.([44])**explaining it to patients is tricky**.' (10)	
		'I think **I wouldn't have a problem with that [a black box model**]… **Unless I had a very statistical patient that was coming in**, that probably has more math than I did.' (08)	

LVAD, left ventricular assist device.

Table 4 Relational considerations for trust

	Physicians	Patients
Peer/expert—reviewed		
	'What's important is if it's based on published results, et cetera, and in good patient population.' (14)	**'I trust it pretty much because I did ask how long have they [the medical team] been doing LVAD and I was surprised to find out that this had been going on longer than I thought it was** and pretty much I said that **they should have enough information right now to tell you what's going to happen** and depending on what stage or condition your heart failure is.' (08)
	'For me, **if it's been validated and published in a peer reviewed journal, I probably would trust it.'** (06)	'It could be accurate. **I would say reluctantly, yes, accurate**… Of course, **you make your trust [based on] the knowledge that you get—the knowledge of the people like my doctors** that have done the study on me individually, rather than on the broad, most people type. **It would be that reason I would probably trust their [my doctors] developing [using] the statistics**.' (17)
	'Probably a paper or the idea that it's [the algorithmic model is] reputed, [if] **it has a good reputation and is well thought of among people that you care about, colleagues whose opinions you care about.'** (07)	

LVAD, left ventricular assist device.

model rather than on the functional nature of the algorithm itself, respectively.

Notably, these concepts do not map perfectly onto existing terms in the 'explainable AI' domain. For example, 'explainability' typically refers to the ability to identify and understand why and how a conclusion was reached by an AI/ML system. Combi *et al*, following a long line of other thinkers in the field,[20–22] distinguish this concept from 'interpretability', which they define as 'the degree to which a user can intuit the cause of a decision (and) to which a human can consistently predict a model's results, based on her experience with the application'.[23] Others[24 25] refer to the latter point as 'reliability'; and the enduring challenges of defining and redefining these terms suggests significant conceptual overlap. Neither of these conceptions of 'interpretability' and 'reliability' clearly addresses the distinction that physicians and patients from our study make in their preferences to know more about data *sources* versus data *processing*. An important implication is that developers seeking to address user trust criteria should be prepared to provide information about data sources in addition to—or potentially in place of—more technical, functional and process-oriented explanations. This will require not only transparency but also some level of expertise in communicating often technical information about data sources (eg, statistical composition of training populations, settings, etc) in user-interpretable ways. This finding enriches emerging understandings in the AI ethics literature around the need for explainability and suggests that explanations of certain—but maybe not all—aspects of AI/ML may be especially important for user trust.

As an extension of source explainability criteria, some patients and physicians rationalised that they could never fully trust *any* predictive model (as a matter of personal belief) because they will never be able to account for all factors influencing outcomes, that is, 'the long tail' of hard-to-anticipate factors. Hypothetically speaking, if perfectly comprehensive data sets were available, algorithmic models would likely demonstrate vastly improved performance compared with humans, as demonstrated by recent advancements in large language models (LLMs), whose ingestion of enormous and topically diverse data sets (eg, large swathes of the internet) permit LLMs like Chat GPT-4 to respond to informational requests with unprecedented content expertise, reasoning and precision.[26] While these systems remain imperfect predictors and are susceptible to artificial 'hallucinations' and other phenomena that are still not well understood,[27] they raise a critical question for human trust in AI/ML: if these models are ever able to ingest *all* available information, will we trust them almost entirely? For the first time, we can pursue this question empirically, given the pace at which AI/ML systems are ingesting and learning from ever greater data sets. As they continue to learn more and more, will we (or should we) continue to recalibrate our trust in them?

Ethical and practical significance of appropriate trust calibration

Until (and whether) users ever have near perfect levels of trust in AI/ML systems, evaluations of trust may be better conceptualised not in terms of absolutes (trust vs distrust) but a continuum (ie, how *much* to trust). Indeed, most physicians and patients in our study talked about how much to trust an algorithm's estimates, reflecting an effort to calibrate trust according to the level of trust one 'should' have in the algorithm. The answer to this normative question is not straightforward. Furthermore, participants did not meaningfully distinguish between actual and normative trust considerations, in other words, how much they *do* versus *should* trust algorithms. However, these are two separate questions that deserve further exploration. Trust calibration should depend on a particular algorithm's performance capacities, the nature of its training data sets and the intended uses of a particular algorithm.[8 28–31] As such, there is not likely to be a 'one size fits all' level at which trust in AI systems should be considered 'appropriately' calibrated. While participants' responses suggest an intention to calibrate trust according to performance capacities and intended uses, whether they do calibrate trust in such a way remains an open empirical question. Further observational or experimental research is also needed into which conditions (interactional, environmental, dispositional) foster appropriate levels of trust. Ensuring a proper level of trust calibration has both practical and ethical implications, given that over-trust or under-trust can lead to over-reliance or under-reliance on an algorithmic tool in practice. In healthcare, missing this mark could have significant implications for patient safety and health outcomes. For this reason, we argue that determining the proper level of trust calibration for a given algorithm or AI tool should be an ethical imperative and an integral part of algorithmic development and validation. However, this analysis will necessarily be contextual.[4 32] We also argue that stated trust (and considerations for trust) should be examined separately from reliance behaviours, as the latter may not map in

expected ways onto the former. This is because multiple factors can influence reliance independent of trust. For example, organisation, cultural and environmental factors such as workload and need for multitasking (eg, in high-pressure or fast-paced environments) can increase reliance on machines by necessity, even if trust is low.[8]

Determining the 'right' level of trust calibration should begin with gauging diverse stakeholders' perspectives for trustworthiness (as we did here) to understand the range of relevant trust considerations. Developers of algorithmic tools and data sets can use this information to clearly outline the degree to which an algorithm aligns with user-defined and consensus trust criteria. Other consensus criteria from the scientific community, such as those identified by the National Institute of Standards and Technology (NIST), should also be addressed. Taken together, conveying the degree to which an algorithm meets stakeholder-specific and expert-identified criteria for trustworthiness will help users to effectively engage in trust calibration and inform decisions about how much stock to put in AI tools. Similar to how Gebru *et al*[33] argue that datasheets describing the creation and use of data sets can help to inform decisions about using a data set, we argue that providing relevant information about an AI tool's development can help end-users appropriately gauge how much to trust (ie, rely on) a tool. Gerke's[34] recent elaboration of AI/ML labels offers concrete examples of what type of information should be considered for inclusion, ranging from information about a tool's intended uses (eg, risk stratification of patients); technical composition of training data sets (eg, statistical comparison of data samples to larger population characteristics); data processing approaches (including classification cut-offs; use of proxies; etc); validation settings and populations; alternative data sets and rationale for their exclusion, among others.

Rejecting trust calibration: the role of personal belief and subjective experience

A minority of our results suggest that some stakeholders are not willing to calibrate their trust, as they have already made up their minds to distrust, citing subjective perspectives anchored in personal experience, emotional dispositions or worldviews that cannot be explicitly explained or elaborated on. For example, a minority of patients who expressed a high distrust in predictive algorithms said that they preferred to make clinical decisions based more on hope/optimism and faith than on statistics. Some respondents pointed to the role of faith, hope and optimism in shaping survival. This belief in the causal superiority of immeasurable, incomputable phenomena invalidates the conclusions of an AI model whose very nature relies on measurement (ie, weights reflecting direction and magnitude) and computation. Because such beliefs do not lend themselves to disproof, trust cannot be easily calibrated towards a normative goal, that is, how much the algorithm 'should' be trusted in a given circumstance.

In another instance, a patient explained his distrust in AI estimates based on his own experiences of negative outcomes that were not anticipated by his clinical team. This patient developed a worldview that predictions, broadly speaking, can never be fully trusted. Rather than reflecting a dispassionate appraisal of an algorithms' specific performance metrics or data sources, this patient's trust was tethered to a more sweeping worldview, confirming a now-established literature that suggests trust and uptake of emerging technologies depend to a large extent on broader attitudes towards science and technology. These perspectives may in turn be linked to cultural beliefs and/or historical injustices related to scientific research that have differentially impacted Black and minority populations. A growing awareness in recent years of racial and gender bias among certain AI tools has understandably rekindled some of this scepticism and distrust.[35 36]

Towards responsible and rational trust calibration—shaping trust

Our exploration of stakeholders' trust criteria offers an example of how to build a baseline understanding of what types of information stakeholders require to appropriately calibrate their trust in a predictive algorithm. However, an even more revelatory indicator of how end-users are likely to employ these tools in practice would be to *observe* trust-related behaviours (eg, frequency of use; evidence of over-reliance/under-reliance) in real-world settings. Researchers have only just begun to engage in such controlled experiments; initial findings demonstrate that reliance on algorithms (and their integration into clinical decision-making) may be influenced by a variety of factors, including task or domain expertise,[9 37] as well as the ability to recognise bias in outputs (eg, consistent over- or under-estimation) and the (separate) ability to correctly address this bias in one's decision-making.[38] Other scholars such as Babic *et al*[39 40] have called for examining human factors and other outcomes of using AI/ML in real world settings through the use of well-designed clinical trials. Similarly, Gerke *et al*[32] have argued that algorithmic tools should be prospectively tested to understand their performance as well as human factors impacting their use across a variety of procedural contexts that mirror intended use settings and human-AI interactions. Such an approach requires that researchers both anticipate and measure a wide variety of potential factors (institutional requirements; professional expectations and conventions; workflow arrangements) that might shape users' behaviours. An implication is that developers and implementors can become more aware — and transparently share this knowledge with end users — of which factors and features of use settings and end-users may significantly influence use patterns and system results.

An additional consideration invoked by our findings is the fact that most stakeholders cited verifiable criteria for trust which implies the possibility of *shaping* users' trust in AI/ML technologies. This modest discovery has rather large ethical implications because it raises questions about 1) whether and how much developers and implementers should be trying to influence (ie, nudge) users' trust in particular ways; and if so, 2) towards what ends (ie, increased uptake?). One perspective is that we (researchers; developers; implementers; etc.) should remain neutral and merely present users with objective datasheets that would allow them to make their own, informed decisions. Another perspective, and one we have explored elsewhere, is that developers and implementors may consider interface design approaches with potential to encourage critical reflection about how to calibrate trust when interacting with AI/ML systems.[41] This logic could be extended to try to nudge[42] users towards greater (or less) trust or reliance, as appropriate. Further empirical research is needed to explore practical and ethical impacts of such approaches.

LIMITATIONS

A major limitation to this exploratory study is that we asked respondents to comment from a hypothetical perspective; it is challenging to discern whether respondents' viewpoints reflect *ideal* considerations for trust or how respondents might engage in actual, real-world decisions and behaviours (eg, reliance) in

relation to AI/ML systems. As AI ethics scholars have pointed out,[32] observational and experimental research is needed to test out user behaviours and attitudes in real world settings and across a range of use contexts in order to observe and understand the degree to which users rely on them in decision making.

CONCLUSION

The effort to calibrate trust according to system performance (accuracy and data source validity --epistemic trust) as well as reputation (relational trust) reflects a rational approach to managing our relationship with AI. These criteria, which we observed to be prevalent among physicians and patients alike, reflect that stakeholders from our study share an empirical approach to evaluating the trustworthiness of predictive algorithms for healthcare. Further, that evaluations for trust focused primarily on the nature, integrity and relevance of training data rather than on critical appraisals of the algorithms themselves suggests a need to distinguish 'source explainability' from 'functional explainability'. Priorities for future research should include ascertaining who will do the work of identifying and communicating the strength and limitations of AI/ML systems, empirically observing how trust-relevant information is received and acted on across diverse applications, settings and populations, and what role developers, clinicians and patients and other stakeholders can each play in ensuring that human-AI interactions reflect appropriate and responsible trust calibration.

Twitter Kristin Kostick-Quenet @kkostick

Contributors KK-Q wrote the initial draft that was reviewed by all other authors, who contributed additions and modifications. JB-B, KK-Q and BHL collaboratively developed the codebook. Interviews were conducted by BHL and MH. KK-Q, BHL, MH and JS contributed to coding. All authors reviewed and approved this manuscript and KK-Q is the guarantor of this work.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The study was approved by the Institutional Review Board at Baylor College of Medicine (H-48537). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iD

Kristin Kostick-Quenet http://orcid.org/0000-0003-2510-0174

REFERENCES

1 Laying down Harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts. 2021.
2 Schwartz R, Vassilev A, Greene K, *et al*. Towards a standard for identifying and managing bias in artificial intelligence. *NIST Special Publication* 2022;1270:1–77.
3 Council US-ETaT. TTC joint roadmap on evaluation and measurement tools for trustworthy AI and risk management. 2022.
4 Varshney KR. Trustworthy machine learning. Chappaqua, NY, 2021.
5 Ajzen I, Fishbein M. *Understanding Attitudes and Predicting Social Behavior*. NJ: Prentice-Hall: Englewood Cliffs, 1980.
6 Biros DP, Daly M, Gunsch G. The influence of task load and automation trust on deception detection. *Group Decision and Negotiation* 2004;13:173–89.
7 Daly MA. Task load and automation use in an uncertain environment. 2002.
8 Lee JD, See KA. Trust in automation: designing for appropriate reliance. *Hum Factors* 2004;46:50–80.
9 Gaube S, Suresh H, Raue M, *et al*. Do as AI say: susceptibility in deployment of clinical decision-AIDS. *NPJ Digit Med* 2021;4:31.
10 Kostick KM, Minard CG, Wilhelms LA, *et al*. Development and validation of a patient-centered knowledge scale for left ventricular assist device placement. *J Heart Lung Transplant* 2016;35:768–76.
11 Blumenthal-Barby JS, Kostick KM, Delgado ED, *et al*. Assessment of patients' and Caregivers' informational and decisional needs for left ventricular assist device placement: implications for informed consent and shared decision-making. *J Heart Lung Transplant* 2015;34:1182–9.
12 Kostick KM, Bruce CR, Minard CG, *et al*. A Multisite randomized controlled trial of a patient-centered ventricular assist device decision aid (VADDA trial). *Journal of Cardiac Failure* 2018;24:661–71.
13 Software V. *VERBI Software Berlin*. 2019.
14 Boyatzis RE. *Transforming qualitative information: Thematic analysis and code development*. sage, 1998.
15 Rempel JK, Holmes JG, Zanna MP. Trust in close relationships. *Journal of Personality and Social Psychology* 1985;49:95–112.
16 Policy. *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American*. The White House, 2022.
17 Bauer K, von Zahn M, Hinz O. Expl (AI) Ned: the impact of Explainable artificial intelligence on users' information processing. *Information Systems Research* 2023.
18 Bussone A, Stumpf S, O'Sullivan D. The role of explanations on trust and reliance in clinical decision support systems. 2015 International Conference on Healthcare Informatics (ICHI); Dallas, TX, USA.
19 Erlei A, Nekdem F, Meub L, *et al*. *HCOMP* 2020;8:43–52.
20 Gilpin LH, Bau D, Yuan BZ, *et al*. Explaining explanations: an overview of Interpretability of machine learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA); Turin, Italy.
21 Joyce DW, Kormilitzin A, Smith KA, *et al*. Explainable artificial intelligence for mental health through transparency and Interpretability for Understandability. *NPJ Digit Med* 2023;6:6.
22 Mittelstadt B, Russell C, Wachter S. Explaining explanations in AI. Mittelstadt B, Russell C, Wachter S, eds. FAT* '19; Atlanta GA USA.New York, NY, USA, January 29, 2019
23 Combi C, Amico B, Bellazzi R, *et al*. A manifesto on Explainability for artificial intelligence in medicine. *Artif Intell Med* 2022;133:102423.
24 Ryan M. In AI we trust: ethics, artificial intelligence, and reliability. *Sci Eng Ethics* 2020;26:2749–67.
25 Shneiderman B. Human-centered artificial intelligence: reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 2020;36:495–504.
26 OpenAI. GPT-4 technical report. 2023.
27 Alkaissi H, McFarlane SI. Artificial hallucinations in Chatgpt: implications in scientific writing. *Cureus* 2023;15:e35179.
28 Chong L, Zhang G, Goucher-Lambert K, *et al*. Human confidence in artificial intelligence and in themselves: the evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior* 2022;127:107018.
29 Glikson E, Woolley AW. Human trust in artificial intelligence: review of empirical research. *ANNALS* 2020;14:627–60.
30 Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust. *Hum Factors* 2015;57:407–34.
31 Jacovi A, Marasović A, Miller T. Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. Goldberg Y, ed. Proceedings of the 2021 ACM conference on fairness, accountability, and transparency; 2021
32 Gerke S, Babic B, Evgeniou T, *et al*. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *NPJ Digit Med* 2020;3:53.
33 Gebru T, Morgenstern J, Vecchione B, *et al*. Datasheets for Datasets. *Commun ACM* 2021;64:86–92.
34 Gerke S. "nutrition facts labels" for artificial intelligence/machine learning-based medical devices-the urgent need for labeling standards". *Geo Wash L Rev* 2023;91:79.
35 Kostick-Quenet KM, Cohen IG, Gerke S, *et al*. Mitigating racial bias in machine learning. *J Law Med Ethics* 2022;50:92–100.
36 Raji ID, Gebru T, Mitchell M, *et al*. Saving face: investigating the ethical concerns of facial recognition auditing. Denton E, ed. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society; 2020
37 Dijkstra JJ, Liebrand WBG, Timminga E. Persuasiveness of expert systems. *Behaviour & Information Technology* 1998;17:155–63.
38 Biermann J, Horton JJ, Walter J. n.d. Algorithmic advice as a credence good. *SSRN Journal*;2022:22–071.
39 Babic B, Gerke S, Evgeniou T, *et al*. Algorithms on regulatory Lockdown in medicine. *Science* 2019;366:1202–4.
40 Babic B, Gerke S, Evgeniou T, *et al*. Beware explanations from AI in health care. *Science* 2021;373:284–6.
41 Kostick-Quenet KM, Gerke S. AI in the hands of imperfect users. *NPJ Digit Med* 2022;5:197.
42 Kostick KM, Trejo M, Volk RJ, *et al*. Using Nudges to enhance Clinicians' implementation of shared decision making with patient decision AIDS. *MDM Policy & Practice* 2020;5:238146832091590.
43 Thenganatt MA, Jankovic J. Treatment of dystonia. *Neurotherapeutics* 2014;11:139–52.
44 Butler JK, Cantrell RS. A behavioral decision theory approach to modeling Dyadic trust in superiors and subordinates. *Psychol Rep* 1984;55:19–28.